Towards automatic detection of prosodic boundaries in spoken French

***Piet Mertens*° & *Anne Catherine Simon*°°**

Piet.Mertens@arts.kuleuven.be, Anne-Catherine.Simon@uclouvain.be
°Linguistics Department, University of Leuven (KU Leuven), Belgium
°°Institut Langage & Communication / Valibel, Université catholique de Louvain, Belgium

Abstract

This paper describes a tool for automatic detection of prosodic boundaries (PBs) in French, and evaluates it on a 12 min. speech corpus, using a reference annotation prepared by a trained phonetician.

The phonetic alignment of the corpus is used to identify the nucleus (as the local peak of intensity) within the rhyme of each syllable. After pitch stylization, the following prosodic properties are computed for each syllable: pause presence, syllable lengthening, intra-syllabic pitch movement, and pitch prominence. These features are combined in detection rules. No training corpus is required.

Perceived PBs of 4 strengths (major, intermediate, minor, no boundary) were annotated by an expert, for a test corpus of 8 samples of continuous speech.

The match between reference PBs and detected PBs was evaluated as a classification task. In addition, the contribution of each prosodic feature to boundary detection was estimated.

1. Introduction

One of the functions of prosody is to divide speech into chunks of one or more words. The term *prosodic boundary* (PB) is used in two ways, to refer either to the limits of these chunks, or to speech properties that mark these limits. According to Ladd (2008: 288), PBs are remarkably difficult to define and to identify consistently and as a result, there is often disagreement about whether a PB is or is not present.

Prosodic units and PBs are relevant for spoken language research and discourse analysis, where segmentation is central for the analysis of turn construction in conversation (Selting 2000), the prosody-syntax interface (Mertens 2006), discourse relations (Wichmann 2000), the understanding of discourse production and processing (Clifton *et al*. 2006), as well as for speech recognition and text-to-speech synthesis.

The *acoustic and perceptual cues* (Wagner & Watson 2010: 907-910) of PBs are related to the presence of a pause, to duration (pre-boundary lengthening, domain-initial strengthening), pitch (pitch movement, pitch discontinuity at the PB, including declination line reset, and the relative height of pitch accents), to intensity, and phonation type (creak).

In prosody research PBs are characterized in various ways (Wagner & Watson 2010: 911). The *qualitative* view assumes a small set of boundary *types*, associated with the corresponding *hierarchical prosodic units* (such as phonological word, intonational phrase), or with functionally defined *boundary strengths* (non-terminal vs. terminal PB, or continuation vs. final PB). In a second, *quantitative* view, the strength of a PB is expressed *relative* to the strength of boundaries occurring earlier in the utterance, without reference to prosodic units or functions. In a third view, boundary strength is approached as a *perceptual* notion in its own right, without reference to prosodic structure, and which can be measured directly in perception experiments with untrained listeners (Pijper & Sanderman 1995). In this paper a PB is defined as a perceived separation between two chunks of speech, marked by prosodic means such as pause, pitch movement, pitch discontinuities and lengthening.

In free stress languages the stressed

syllable and the syllable preceding the PB are often dissociated. As a result, acoustic correlates of PBs may be associated with the pre-boundary syllable(s), or with events (pause, pitch discontinuity…) at the PB itself. Instead, in some fixed stress languages, stress usually occurs at a PB, in which case both stress and boundary features are associated with the same syllable. This is the case in French, where word stress is on the final syllable of an intonation unit (Mertens 1993; Di Cristo 1999).

The variety of PB definitions raises the question whether PBs may be *identified consistently* by human annotators. In manual labelling by trained listeners, based on perception, judges may rely on their ideas about prosodic organization and use the segmental information to identify syntactic structure (Campbell, 1993: 10; Pijper & Sanderman 1995: 2038). However, perception experiments (Pijper & Sanderman 1995) suggest that untrained listeners can give reliable judgments of PB strength, even when the lexical contents of the utterances is made unrecognizable ("delexicalised").

How many *levels of boundary strength* should be distinguished? In perception experiments where listeners are asked to rate PBs on a 10-point scale, the number of resulting PB categories is usually lower (Pijper & Sanderman 1995). Grover *et al.* (1997) suggests that boundary strength is transcribed more consistently using a 4-value scale than using a scale with a many values. But often the number depends on theoretical assumptions about PBs. Phonological models of French intonation (e.g. Martin 1978, Rossi 1999, Mertens 2006) often imply three or more levels of boundaries (cf. minor and major continuation, and terminal contour of Delattre 1966). Annotated corpora commonly use two or three levels: the C-ORAL-ROM corpus annotation (Moneglia *et al*. 2005) distinguishes terminal and non-terminal PBs.

Automatic detection of PBs relies on their acoustic cues (cf. supra), although the approaches and algorithms for the actual detection differ considerably (Ostendorf 2000). The experiments of Pijper & Sanderman (1995), using judgments by untrained listeners, suggest that, for Dutch read speech, the most important cues of PBs are the presence of a pause (> 200ms), the melodic discontinuity (drop of pitch during the silent PB), the declination line reset and pre-boundary lengthening. The relation between PBs and temporal organisation (duration of phonetic segments in syllable onset, nucleus and coda) is studied in detail in Campbell (1993, 2000). There have been efforts to combine into a single algorithm the detection of pauses, lengthening, F0 variation, prominence, in order to segment spoken French into major prosodic units (Lacheret-Dujour & Victorri 2002), also using lexical information.

2. The system for PB detection in French

Our strategy strongly relies on the prosodic structure of French, in which the last syllable of the intonation unit may carry particular pitch movements, may be prominent for pitch or duration (lengthening), and followed by a pause.

2.1. Prosodic features measured

The following properties are measured: pause presence, syllable lengthening, intra-syllabic pitch movement, and pitch prominence.

A syllable is *prominent for some prosodic attribute* (duration, pitch) when it stands out from its context due to a local difference for that attribute (Mertens 1991). Prominence may be quantified as the value at the target syllable, divided by the mean value in the context. Depending on the number of syllables in the left and right context, the context window may be symmetrical or asymmetrical, fixed or dynamic; in the latter case, window length depends upon the properties of the syllables in the context.

The *phonetic alignment* provides the syllables, vowels and rhymes. The *syllabic nucleus* is determined as the voiced part of the rhyme (= vowel + coda), located around the intensity peak of the vowel, for which the intensity drop stays below some threshold and provided it does not include pitch discontinuities.

A *pause* is detected when the interval between successive nuclei exceeds 200ms.

Hesitations ("euh" vowels in French) affect the estimation of lengthening and pitch prominence, more generally when they appear in the context used for measuring prominence. A hesitation is detected when a syllable is labelled [œ], [ø] or [ə], has a duration of at least 350 ms and a pitch which is level to slightly falling (> -3 ST).

Syllable lengthening is measured as syllable duration prominence (*SDP* = syllable duration / mean syllable duration in context window), for an asymmetrical and dynamic context window of at most 2 syllables to the left and 1 to the right.

The following problems were encountered. (1) Hesitations are most often considerably longer than other syllables, affecting syllable length measurement. (2) Pauses act as perceptual boundaries for the context window. Therefore, the context window is truncated at a pause, and its width is adapted dynamically (max. 500ms for each side). To avoid artefacts due to small context size, *SDP* is set to 1 (hence, no lengthening) when the context contains only 2 syllables. (3) At high speech rate, intrinsic duration of speech sounds largely affects sound duration. To avoid this, *SDP* is set to 1 for nuclei of 40 ms or shorter.

Pitch prominence is measured as prominence of the mean pitch value (the mean F0 within the syllabic nucleus), for a context size of 2 syllables to the left and 1 syllables to the right, using a dynamic context width.

The values for intra-syllabic *pitch rise* and *pitch fall* indicate the cumulated positive, resp. negative, pitch intervals within the syllabic nucleus.

2.2. Rules for PB assignment

The rules for PB assignment given below are based on empirical observation of corpus data annotated by a phonetician. They are similar to those of other studies (e.g. Lacheret & Victorri 2002).

1. Do not assign a PB to hesitation syllables (which are detected automatically).
2. Assign a *major PB* (level 3) in three cases: (a) when duration prominence (SDP) exceeds 3 (i.e. 3 times as long as the context mean, corrected for high speech rate), (b) when the nucleus contains a pitch rise or a pitch fall of 10 ST or more, or (c) when it is followed by a pause of at least 200 ms.
3. Assign an *intermediate PB* (level 2) in three cases: (a) when SDP exceeds 2 (corrected for high speech rate), (b) when the pitch rise in the nucleus is at least 4 ST, or (c) when pitch prominence is 5 ST or more.
4. Assign a *minor boundary* (level 1) when SDP exceeds 1.5, provided nucleus duration is at least 40 ms.

3. Evaluation

3.1 Speech material

The test corpus contains 8 speech samples[1] of approx. 100s, by 9 speakers, male and female, with a total duration of 737s (3029 syllables). Four samples consists of unprepared speech (radio-interviews, conversations), the other four of read-aloud speech (radio-news, conference presentations). A

[1] All samples, except the directions request, are taken from the Valibel Speech Database (Dister *et al.* 2009) and illustrate a (standard) Belgian variety of French. Samples under the category "academic" represent academic discourse at official occasions. Radio-news and radio-interview are broadcasts from the national radio programs. Interview comes from sociolinguistic investigation about standard usages of French. Directions request comes from a M. Avanzi corpus collected in France, and available in the C-Prom project (Avanzi *et al.* 2010). Finally, everyday conversation involves two close friends self-recorded at home.

validated phonetic alignment was prepared by the authors. Subcorpus A contains all 3029 syllables; in subcorpus B hesitations (detected or marked in the corpus annotation) and syllables without a detected syllabic nucleus (either because it was unvoiced, too short or contained pitch discontinuities), are discarded, resulting in a set of 2625 syllables.

3.2. Reference annotation

Manual labelling of the speech material was carried out by a trained phonetician, who is a native speaker of French. The annotator listened to sound fragments of 4 to 8s, played 3 times. For each word-final syllable (i.e. a potential stress), the annotator assigned one out of four boundary levels, either 0 (no PB), 1 (minor PB), 2 (intermediate PB) or 3 (major PB). Syllables with emphatic initial stress (ES) and hesitations were also identified by the annotator, but treated as "no PB" in the evaluation described here.

3.3. Results

The automatic PB detection was evaluated as a classification task, mapping *observed* categories to *actual* ones. In this case, the observed category is the automatically detected PB and the *actual* category is provided by the reference labelling of the phonetician.

For *minor PBs* (level 1) very poor results are obtained. In French, pitch contours are anchored at the final syllable of an intonation group, which coincides with the last syllable of a syntactic constituent, and PBs are very likely at the end of a syntactic constituent. As a result PB perception may be biased by segmental information about syntactic structure. This holds for PBs of all levels, of course, but for PBs of level 2 and 3 acoustic cues are usually present. In the evaluation, minor PBs were removed (i.e. replaced by "no PB") from the reference annotation and the detected PBs.

The confusion matrix for subcorpus B is shown in table 1. The detection of major PBs (level 3) is rather good, whereas that of intermediate PBs (level 2) is poor.

	observed			
actual	0	2	3	
0	2041	112	80	2233
2	69	81	24	174
3	11	16	191	218
	2121	209	295	2625

Table 1. Confusion matrix for subcorpus B. (observed = detected PB; actual = reference PB)

The system detects noticeably more level 2 PBs than the human expert (209 vs. 174). This is partly explained by the fact that the human annotator distinguishes between final stress (followed by a PB) and emphatic initial stress ("ES", treated as "no PB" in the evaluation data), whereas the system does not detect ES as such, and as a result many syllables carrying ES will be detected as a level 2 PB (but not as a level 3 PB, since ES is not followed by a pause).

80 syllables without a PB were detected as major PBs. This is explained by the fact that the last syllable of an utterance is often devoiced, in which case its nucleus is not detected, and no PB will be detected either. Also, the skipped syllable may be interpreted as a pause, resulting in a major PB being detected at the preceding syllable.

	PB	N	%	Prec.	Rec.	Accur.	F
A	0	2586	85.4	94.8	91.7	88.6	93.2
	2	182	6.0	37.7	44.5	92.2	40.8
	3	259	8.6	61.4	73.7	93.8	67.0
A'	0	2586	85.4	94.8	91.4	88.4	93.1
	2	182	6.0	37.6	45.1	92.2	41.0
	3	259	8.6	59.9	73.7	93.5	66.1
B	0	2233	85.1	95.7	93.8	91.1	94.7
	2	174	6.6	38.6	44.8	91.6	41.5
	3	218	8.3	79.0	84.4	96.8	81.6

Table 2. Classification results for the automatic detection of PBs using 3 boundary levels (0, 2 and 3) for subcorpora A, A' (see text) and B. (Data set does not include level 1 PBs.) N = count.

Table 2 shows precision, recall, accuracy and F-measure for each PB class (0, 2, 3), as well as the number (N) and percentage (%)

of elements in each class, for subcorpora A and B. The elimination of hesitations and syllables without a detected nucleus – in subcorpus B – slightly improves recall and precision, in particular for major PBs: 84.4% of the actual major PBs are indeed detected as major PBs (recall) and 79% of the detected major PBs are actual major PBs (precision). Syllables without a PB are detected with a precision of 95.7% and a recall of 93.8%.

Finally, results for "A'" correspond to subcorpus A, when decision rule 1 is disabled, i.e. when hesitations are treated in the same way as other syllables. The small difference between A and A' shows the impact of rule 1 is negligible.

Table 3 shows the contribution of individual prosodic cues to the detection of *actual* PBs of level 2 and 3, as the percentage of boundaries for which a given cue was present. In subcorpus B, for *major* PBs, pause is the most effective cue (85.8%), followed by lengthening (11%). For *intermediate* PBs the most important cues are lengthening (24.1%) and pitch prominence (24.1%), whereas the contribution of pause drops to 9.8%.

PB	N	P	R	F	r	T	L2	L3
2	174	9.8	0.6	1.7	6.9	24.1	24.1	2.3
3	218	85.8	5.5	1.4	4.6	6.0	1.4	11.0

Table 3. Contribution (percentage present) of prosodic cues to PB detection for PB levels 2 and 3, in subcorpus B. N=number of syllables, P=pause, R=large rise (≥10ST), F= large fall (≤10ST), r=small rise (≥4ST) , T=pitch prominence (≥5ST), L2=lengthening ≥2, L3=lengthening ≥3 .

Using pause as the *only* cue results in the detection of 280 *major* PBs, against 295 when *all* cues are used. Note that the number of observed (detected) PBs exceeds the number (218) of actual major PBs. Table 4 shows the classification results for *major* PBs only, when level 2 PBs are treated as "no PB". It shows the use of cues other than pause improves recall form 85.8% to 87.6%. The scores of tables 2 and 4 should not be compared, since they represent distinct classification tasks: 3 classes (0, 2, 3) for table 2, against 2 classes (0, 3) for table 4.

cues used	N actual	N obs.	Prec.	Rec.
pause only	218	280	66.8	85.8
all cues	218	295	64.7	87.6

Table 4. Classification results for subcorpus B for the automatic detection of major PBs using either pause only or all prosodic criteria.

4. Discussion

Common causes of errors may be identified. The first, syllable devoicing, occurs when one or more syllables at the end of an utterance are pronounced with gradual or complete devoicing, possibly with creak, for instance when sub-glottal pressure decreases or when the pitch drops to the bottom of the pitch range. Such unvoiced syllables will not be recognized as syllables by the algorithm, either due to low intensity, lack of voicing, or octave jumps typical of creak.

In French, the syllable with final stress may sometimes be followed by a schwa, detected as a separate syllable or even as a hesitation. This will affect the location of the detected PB.

In a third type of error a PB is detected at an initial emphatic stress. Currently the system is unable to distinguish the two types of stress found in French.

The fourth type of error concerns hesitations. The perception of hesitation is a complex phenomenon, which implies prosodic cues (lengthened syllables, flat or slightly falling pitch contour) but also syntactic phenomena (intra-phrase silent pauses, repetitions and false starts, cf. Duez 2001). Only 25% of the hesitations in the reference annotation are detected by the system.

5. Conclusion

Speech corpora are largely available today, but still require more reliable tools for annotation tasks, especially for prosodic annotation. In this contribution, we propose

an automatic tool for PB detection in spoken French. Detection is purely acoustic: it is based on a detection of syllable prominence (pitch movement or pitch peak, lengthening) in a local context, and pause. One out of three boundary levels is assigned, depending on the characteristics of the syllable. Good results are obtained for detection of major PBs and acceptable results for the detection of intermediate PBs.

Analysis of the results shows that the most important cue for *major* PBs is pause (85.8%), followed by lengthening (11%), whereas for *intermediate* PBs, the most effective cues are lengthening (24.1%) and pitch prominence (24.1%).

The analysis of frequent errors suggests improvements of the algorithm. First, access to lexical information (syllable position within the word, detection of repetition or hesitation particles) would help in interpreting prosodic variation (like syllable lengthening) that may fulfil very diverse functions, according to its location. Second, a better description of the acoustic properties of emphatic initial stress (ES) in French might provide us with tools for distinguishing final vs. initial stress.

References

Avanzi, M., Simon, A.C., Goldman, J.-P. & A. Auchlin. (2010). C-PROM. Un corpus de français parlé annoté pour l'étude des proéminences. *Actes des 23èmes JEP* (Mons, 25-28 mai 2010).

Campbell, W.N. (1993). Automatic detection of prosodic boundaries in speech. *Speech Communication* 13, pp.343-354.

Campbell, W.N. (2000). Timing in Speech: A Multi-Level Process. Horne, Merle (ed.) *Prosody: Theory and Experiment.* Dordrecht, Kluwer.

Clifton, C., Carlson, K. & Frazier, L. (2006). Tracking the what and why of speakers' choices: prosodic boundaries and the length of constituents. *Psychonomic Bulletin & Review* 13:5,pp. 854-61.

Delattre, P. (1966). Les dix intonations de base du français. *French Review* 40/1, pp.1-14.

de Pijper, J. R. & Sanderman, A. (1995). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *JASA* 96, pp. 2037-2047.

Di Cristo, A. (1999). Vers une modélisation de l'accentuation en français (première partie). *J. of French Language Studies* 9:2, pp. 143-163.

Dister, A., Francard, M., Hambye, Ph. & Simon, A.C. (2009). Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de banque de données textuelles orales VALIBEL (1989-2009). *Cahiers de Linguistique* 33:2, pp. 113-129.

Duez, D. (2001). Signification des hésitations dans la production et la perception de la parole spontanée. *Revue PArole* 17-18-19, pp. 113-137.

Grover, C., Heuft, B., Coile, B. van (1997). The reliability of labeling word prominence and prosodic boundary strength. *Proc. ESCA Workshop on Intonation*, Athens, pp. 165-168.

Lacheret, A. & Victorri, B. (2002). La période intonative comme unité d'analyse pour l'étude du français parlé: modélisation prosodique et enjeux linguistiques, *Verbum* XXIV/1-2, pp. 55-72.

Ladd, D.R. (2008). *Intonational Phonology*. Cambridge University Press, Cambridge.

Martin, Ph. (1978). Questions de phonosyntaxe et de phonosémantique en français. *Lingvisticae Investigationes* 2, pp. 93-126.

Mertens, P. (1991). Local prominence of acoustic and psychoacoustic functions and perceived stress in French. *Proceedings of the 12th International Congress of Phonetic Sciences* 3, pp. 218-221.

Mertens, P. (1993). Intonational grouping, boundaries, and syntactic structure in French. House, D. & P. Touati (eds.), *Proc. ESCA Workshop on Prosody* (Sept. 27-29, 1993, Lund), pp. 156-159.

Mertens, P. (2006). A Predictive Approach to the Analysis of Intonation in Discourse in French. In Kawaguchi, Y.; Fonagy, I.; Moriguchi, T. (eds.), *Prosody and Syntax*. John Benjamins, Amsterdam, pp. 64-101.

Moneglia M., Fabbri M., Quazza S., Panizza A., Danieli M., Garrido J., Swerts, M. (2005). Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM corpus. Moneglia, M. (ed.), *C-ORAL-ROM: integrated reference corpora for spoken Romance language*, Benjamins, pp. 257-276.

Ostendorf, M. (2000). Prosodic Boundary Detection. Horne, Merle (ed.), *Prosody: Theory and Experiment.* Kluwer, Dordrecht.

Rossi, M. (1999). *L'intonation, le système du français: description et modélisation*. Ophrys, Paris-Gap.

Selting, M. (2000). The construction of units in conversational talk. *Lang. in Society* 29, pp. 477-517.

Wagner, M. & Watson, D.G. (2010). Experimental

and theoretical advances in prosody: A review. *Language and Cognitive Processes* 25, pp. 905-945.

Wichmann, A. (2000). *Intonation in Text and Discourse. Beginnings, middles and ends*. Harlow, Longman.